PMVC: Promoting Multi-View Consistency for 3D Scene Reconstruction

Chushan Zhang[1,*], Jinguang Tong[1,2,*],Tao Jun Lin[1], Chuong Nguyen[1,2], Hongdong Li[1]
[1]The Australian National University [2]Data61, CSIRO
[*]These authors contributed equally to this work.
{chushan.zhang, jinguang.tong, taojun.lin, hongdong.li}@anu.edu.au
{jinguang.tong, chuong.nguyen}@data61.csiro.au

Abstract

Reconstructing the geometry of a 3D scene from its multi-view 2D observations has been a central task of 3D computer vision. Recent methods based on neural rendering that use implicit shape representations, such as the neural Signed Distance Function (SDF), have shown impressive performance. However, they fall short in recovering fine details in the scene, especially when employing a multi-layer perceptron (MLP) as the interpolation function for the SDF representation. Per-frame image normal or depth-map prediction have been utilized to tackle this issue, but these learning-based depth/normal predictions are based on a single image frame only, hence overlooking the underlying multiview consistency of the scene, leading to inconsistent erroneous 3D reconstruction. To mitigate this problem, we propose to leverage multi-view deep features computed on the images. In addition, we employ an adaptive sampling strategy that assesses the fidelity of the multi-view image consistency. Our approach outperforms current state-of-the-art methods, delivering an accurate and robust scene representation with particularly enhanced details. The effectiveness of our proposed approach is evaluated by extensive experiments conducted on the ScanNet and Replica datasets, showing superior performance than the current state-of-the-art.

1. Introduction

Reconstructing the 3D mesh of a scene from a sequence of videos or multi-view images captured by moving cameras is crucial for applications like Augmented Reality (AR), Virtual Reality (VR), robotics, and more. Traditional techniques for 3D visual reconstruction, such as Structure from Motion (SfM) or Multi-View Stereo (MVS) algorithms [8, 35, 51], often struggle to recover textureless regions [9], for instance, a blank wall or uniformly colored furniture. This is partly due to the heavy reliance on accurate image feature matching employed by those traditional methods, which fail easily in those textureless regions. Recently, neural rendering-based method using implicit surface representation [29, 42] has emerged as a promising approach for 3D scene reconstruction, often based on the multi-layer perceptron (MLP) network architecture. The basic idea of this method involves using neural networks to represent the geometry and appearance of the scene [25], as opposed to explicit 3D geometry reconstruction.

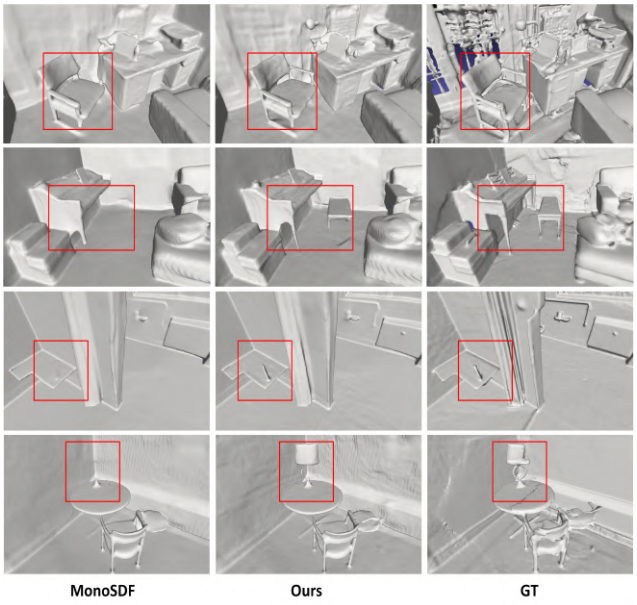


Figure 1. Qualitative comparison of our proposed technique with SOTA [56] and the ground truth. Rectangles highlight the improvements by our technique.

For instance, NeRF [26] employs neural networks to implicitly represent scene properties such as density and color conditioned on spatial coordinates and viewing direction. These properties can then be used to synthesize images from new perspectives. Similar to NeRF, NeuS [48] utilizes an

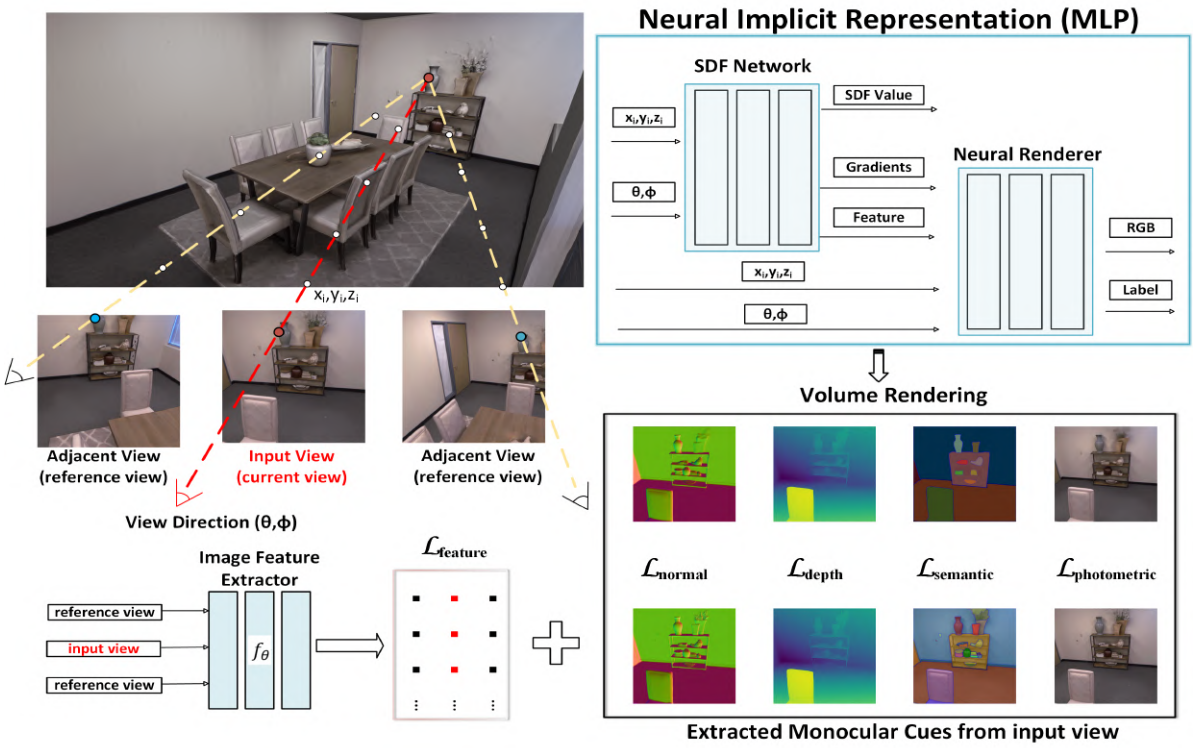


Figure 2. **Overview:** In this research, we utilize a neural feature extractor, introducing intermediate supervision constraints as depicted in Figure 3, and incorporating with supplemental semantic information to enhance multi-view consistency. Specifically, for each ray, we render both the predicted RGB color and our additionally defined semantic labels for optimization. We further compare the features of adjacent views for each ray, expecting similar contributions to the surface from identical points. Furthermore, to achieve higher-quality reconstructions, we adhere to certain geometric priors to assist in optimization.

MLP to model 3D scenes by the representation of signed distance function (SDF). Unlike Mesh or Point Cloud, SDFs implicitly represent points inside or outside the object while maintaining its differentiability.

While implicit neural representation has achieved promising results, it continues to struggle with the effectiveness of reconstructing detailed structures [26, 40, 41]. In order to address this issue and further enhance the quality of reconstruction, researchers have incorporated depth information as additional geometric constraints in their work [1, 6, 50]. While this significantly improves reconstruction quality, the availability of accurate depth information is not always guaranteed. Recent advancements [47, 56] adapt pre-trained models to predict depth and normal maps from monocular color images. These predicted images are then utilized as pseudo-supervision. Despite the potential benefits of these constraints, such methods suffer from the challenge of geometric inconsistency. Specifically, although the predictions for an individual image might exhibit reasonable accuracy, discrepancies arise when these predictions are compared against those generated from a reference viewpoint after projection. This inconsistency poses a dilemma for the neural network as it endeavors to accommodate all the provided supervision, leading to sub-optimal reconstruction results.

In contrast to existing methods that primarily rely on monocular constraints, we propose a new approach to promote multi-view consistency for 3D Scene Reconstruction. To do this, we try to extract multiple levels of information from the color image. This combines the RGB information itself as a fundamental constraint, a feature map as an intermediate constraint, and a predicted semantic map as a high-level constraint. With those multiple constraints, our method addresses global consistency and conducts more accurate 3D reconstructions. Specifically, to address the shortcomings of monocular prediction, we seek a condition that provides a stable prior and satisfies multi-view consistency simultaneously. An important but often overlooked prior is the feature pattern across multiple image views. Traditional CNN training processes rely on inductive learning, where the network convolves each image and identifies different pixel-based features to determine the class to which a given input may belong. To tackle the challenge of reconstructing objects, we contend that the pattern features of points on the same surface should be consistent across multiple frames. By incorporating this constraint into our model, we promote

a supervised learning process that ensures the output satisfies the consistency of pattern features of points on the same surface. Other than introducing multi-level constraints during the training phase to enhance multi-view consistency, it is important to note that, due to our use of the SDF function for representing continuous surfaces and our random sampling method based on other baseline techniques, smaller objects in scenes might be under-sampled during training. As a result, the model may overlook these smaller objects in its quest for a global optimum. To address this issue, we use dynamic sampling of additional edge regions and demonstrate that our method can mitigate the drawbacks of over-smoothing and random sampling. As Figure S8 shows, the proposed approach effectively improves 3D reconstructions that rely on monocular constraints.

Our pipeline yields promising results on real-world scenes using the ScanNet dataset [4], and we discuss our model's performance in detail. Furthermore, we conduct both quantitative and qualitative evaluations using the Replica dataset [44], achieving state-of-the-art performance. In this paper, we incorporate the Neural Renderer model [54]. This work characterizes geometry through the zero-level set of an MLP, offering a more tailored approach for our task.

Our contributions are summarized as follows:

- We propose multi-level constraints for promoting multiview consistency in a neural scene reconstruction task. Our pixel-based and semantic-based priors enforce the model to understand the scene with constraints of local appearance and global semantic consistencies.
- We introduce an adaptive sampling strategy to sample stable pixel correspondences, which provides extra sampling on regions that we are interested in preserving finer details, leading to enhanced performance.
- We evaluate our proposed pipeline on multiple datasets, achieving state-of-the-art quantitative and qualitative results.

2. Related Works

2.1. Traditional multi-view 3D reconstruction

Traditional 3D reconstruction typically commences with feature point matching, followed by generating a sparse point cloud through a SfM process [35,43]. Leveraging this sparse point cloud, a dense 3D reconstruction of the scene can be achieved using an MVS algorithm [8, 37], which incorporates color information and spatial geometric constraints of each pixel. The MVS process generally entails depth map computation, depth map fusion, and surface reconstruction stages [16,20]. Nevertheless, despite improvements brought about by deep learning [13,46,53], this conventional approach still struggles with weakly textured and numerous repetitive regions. Furthermore, the method is constrained by the discrete level of image input; the reconstruction results often exhibit gaps when the distance between each image is substantial. To mitigate these issues, neural network-based implicit reconstruction methods have been developed, providing solutions to many of these challenges.

2.2. Neural implicit 3D Reconstruction

The advent of neural network reconstruction methods can be traced back to the utilization of multilayer perceptrons (MLPs) for scene representation [42]. A groundbreaking study, NeRF and its successors [2, 23, 26, 58], took advantage of the low memory footprint and remarkable representational power of neural networks. The core concept of the NeRF series is to learn a continuous 3D representation of a scene using a fully connected network. The model straightforwardly predicts the density and 3D features of the surface from 3D coordinate inputs. However, NeRF is difficult to extract a high-quality surface since it only learns a volume density field. Building upon this foundational idea, NeuS [48] incorporates insights from DeepSDF [31] to propose a novel approach to volume rendering. In a similar vein of representing scenes through neural networks, IDR [54] reconstructs surfaces by characterizing the geometry as the zero-level set of an MLP, which is considered to be an SDF function. These methods have the ability to accurately delineate an object's surface while maintaining spatial continuity, offering the potential for subsequent surface detail enhancement.

2.3. Incorporating Priors for Reconstruction

Incorporating a prior has been employed to augment the capabilities of the neural network approach. Despite the impressive outcomes demonstrated by neural reconstruction techniques, further enhancements are required to handle edge details and weakly textured regions. As a result, methods exploiting priors have been extensively proposed. Among these, geometric-prior-based approaches, such as those cited in [6, 14], excel at guiding models with geometric information. These methods have demonstrated a significant enhancement in reconstruction results, as indicated by [1]. However, securing accurate geometric data often poses challenges. Consequently, many solutions, like [47, 56], have turned to pseudo-ground-truths, leveraging depth and normal maps from various sources, including monocular predictions. Furthermore, recent research has delved deeper, highlighting the significance of non-geometric priors in this field. For instance, [3, 29, 49, 57] integrates feature information to derive more implicit priors, while studies such as [7, 11, 52, 59] combine semantic infor-

mation in jointly trained models to enhance performance. However, these methods might not fully harness the potential of multi-view reconstruction, potentially leading to suboptimal results in textureless areas or with smaller objects. This limitation often arises from either overlooking the ample information available across multiple views or being restricted by the capabilities of pre-trained models. In contrast, our method emphasizes multi-view consistency, integrating geometric information and merging semantic labels with learnable features across various views. Our proposed approach not only elevates the quality of the reconstruction but also reduces reliance on the proficiency of pre-trained models. Even using a basic model, such as Resnet-18 [12], for our experiments, our methodology outperformed state-of-arts, underscoring its efficacy.

3. Methods

In this paper, we introduce a new method called PMVC, as illustrated in the overview in Figure 2. Our method primarily focuses on indoor scene reconstruction, with RGB images and their camera poses serving as the main inputs to our pipeline. Building upon the previous research [47, 54, 56], we employ pre-trained networks to predict geometric and semantic priors for each frame in the reconstruction process. During the training phase, we introduce a novel multi-view regularization technique to constrain our model by incorporating semantic and feature consistency between a source frame and its reference frame.

In Section 3.1, we review implicit scene representation and the utilization of SDF to represent the volume rendering method. Following that, we discuss how the proposed multi-level constraints can contribute to a better 3D scene reconstruction in Section 3.3. In Section 3.2, we demonstrate our reprojection refinement mechanism to mitigate the side effects of learned monocular priors. Furthermore, we introduce our dynamic sampling strategy in Section 3.4, which facilitates the reconstruction of more significant regions while preserving multi-view consistency. Finally, in Section 3.5, we discuss the formulation of different losses and the implementation of the overall optimization process.

3.1. Implicit Scene Representation

Scene Geometry Representation with SDF. The geometry of a scene can be learned implicitly with a neural Signed Distance Function (SDF) [31], representing its geometric surface as the zero iso-surface decision boundaries of a neural network (MLP). Let SDF be a continuous function f, which outputs the distance s to the nearest surface for each sampled point, such that:

$$f : \mathbb{R}^3 \to \mathbb{R}, \quad \mathbf{x} \mapsto s = \text{SDF}(\mathbf{x}). \tag{1}$$

For any 3D coordinate $\mathbf{x} = (x, y, z)$ mapped with a positional encoding γ [26], its SDF value s can be learned through a single MLP parameterized with learnable parameter ϕ [31]:

$$s = f_\phi(\gamma(\mathbf{x})). \tag{2}$$

Scene representation with Volume Rendering. Differentiable volume rendering techniques, such as NeRF [26], have demonstrated the ability to learn continuous implicit representations for both the geometry and appearance of a scene, simply under the supervision of RGB images. In volume rendering, the rate of occluded light at point $\mathbf{x}$ is expressed with a scalar volumetric function $\sigma(\mathbf{x})$, which denotes the volume density [24]. We follow VolSDF [54] to model the transformation from the learnable SDF f_ϕ to the volume density σ parameterized with learnable parameters $\alpha, \beta > 0$:

$$\sigma(\mathbf{x}) = \alpha \Psi_\beta(f_\phi(\gamma(\mathbf{x}))) \tag{3}$$

where Ψ_β is the Cumulative Distribution Function of a zero mean Laplace distribution with β scale:

$$\Psi_\beta(s) = \begin{cases} \frac{1}{2\beta} \exp\left(\frac{s}{\beta}\right), & \text{if } s \leq 0 \\ \frac{1}{\beta}\left(1 - \frac{1}{2}\exp\left(-\frac{s}{\beta}\right)\right), & \text{if } s > 0. \end{cases} \tag{4}$$

Subsequently, we follow a similar volume rendering formulation to render the appearance of a scene in this work. Let $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ be a ray cast from the optical center $\mathbf{o}$ along the view direction $\mathbf{d}$, we sample M points along the ray with within the bound t_n and t_f. The expected color $\hat{C}(\mathbf{r})$ of camera ray $\mathbf{r}$ is approximated with the following numerical integration:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{M} T_{\mathbf{r}}^i \alpha_{\mathbf{r}}^i \hat{c}_{\mathbf{r}}^i \quad T_{\mathbf{r}}^i = \prod_{j=1}^{i-1}(1-\alpha_{\mathbf{r}}^j) \quad \alpha_{\mathbf{r}}^i = 1-\exp(-\sigma_{\mathbf{r}}^i \delta_{\mathbf{r}}^i), \tag{5}$$

where T_r^i denotes the accumulated transmittance and α_r^i is the α value along ray r from t_n to t_i for each sampled point i, and δ_r^i represents the distance between adjacent samples.

In a similar manner, we can predict the expected depth $\hat{D}(\mathbf{r})$ and surface normal $\hat{N}(\mathbf{r})$ of any surface point the camera ray $\mathbf{r}$ hits:

$$\hat{D}(\mathbf{r}) = \sum_{i=1}^{M} T_{\mathbf{r}}^i \alpha_{\mathbf{r}}^i t^i, \quad \hat{N}(\mathbf{r}) = \sum_{i=1}^{M} T_{\mathbf{r}}^i \alpha_{\mathbf{r}}^i \hat{n}_{\mathbf{r}}^i \tag{6}$$

where $\hat{n}_{\mathbf{r}}^i$ is the analytical normal vector calculated from the gradient of SDF.

3.2. Multi-Level Constraints

The foundation of multi-view 3D reconstruction lies in the principle of multi-view consistency. This principle serves as a cornerstone for both traditional methods, such as SfM and MVS, as well as more recent implicit neural representation approaches. Initially, only photometric information is used as supervision for reconstruction (Eq 9), but

its reliability is compromised by the change in lighting conditions and noises. In our approach, we incorporate multi-level constraints by extracting multi-resolution feature maps for corresponding pixel. This allows us to correlate scene appearance across adjacent views with extract features instead of pixel values. On top of this, an additional learned semantic prior is used as our high-level guidance.

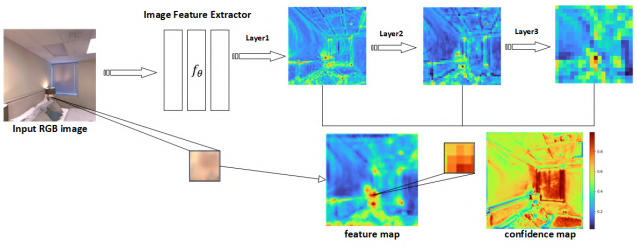


Figure 3. The flowchart illustrates our methodology for generating multiple feature maps using a VGG-19 encoder. Information from each layer is connected via skip-connections, and subsequently, upsampled and merged to create the final feature map and its confidence map. Furthermore, we obtain reprojected features from the interpolated feature maps for incorporating with feature loss in our framework.

Feature Prior In the practical context of multiview reconstruction, we work with the assumption that there is a relatively small motion between neighboring frames. By making this assumption, we are able to formulate constraints on the locally extracted features. To clarify, identical surface points in neighboring frames will be constrained locally in the feature space to maintain their consistency in the image space. Existing research suggests that deep features are typically more stable against illumination changes and motion blur than using photometric intensity alone [19]. Based on this proposition, we have incorporated CNN features as additional intermediate constraints in our approach. Specifically, we use the VGG-19 model as our learnable feature extractor f_θ, which outputs a multi-scale feature map and its corresponding confidence matrix (Figure 3).

For each batch-sampled point, we project it onto its reference view with the given camera poses and predicted depth from our model. Since only unoccluded points should contribute to feature consistency in context of multi-view reconstruction, we select confidence points by evaluating their normal and depth (eq.8). The selection strategy will be detailed in section 3.3. Finally, we weight the filtered points using the corresponding confidence map from our feature extractor, which will be used for the subsequent feature loss.

During our training process, we randomly perturb the location of our reprojected points under uniform distribution to extract negative feature matches. We then include both positive and negative feature matches in the optimization stage as a multi-view feature consistency supervision. This allows us to finetune the feature extractor in our framework for robust feature outputs in the indoor scenes. We anticipate that these features will remain consistent for the same surface point in the scene, such that our model will predict more accurate depth values.

Semantic Prior In the novel-view synthesis task, [59] showed that integrating 2D segmentation can enhance the model's understanding of the scene and yield improved results. We aim to extend this finding by incorporating 2D segmentation into our 3D reconstruction framework. A challenge we wish to address is the absence of ground-truth. To address this issue, we designed a pipeline to generate pseudo ground-truth semantic maps. We first employ GroundingDINO [21] to generate bounding boxes of detected objects. These bounding boxes are used to guide the semantic segmentation process. Then we derive the final pseudo ground-truth semantic maps with Segment-Anything [18], where label information is provided by the detected bounding boxes. We also pre-process the NYU labels [28] by removing certain ambiguous labels to circumvent inconsistent predictions. For additional details on our pseudo ground-truth semantic map generation procedure, please refer to the supplementary materials.

During our experiments, we have observed possible inconsistencies in predicted pixel labels at different views when leveraging the pre-trained model, same phenomenon is discovered in [38]. To address this, we use generated pseudo ground-truth semantic maps in the warm-up stage of model guidance. As the process progresses, we gradually reduce the weights and begin to introduce labels predicted by the model across multiple frames in the later training phase (Eq 11). By further incorporating our learnable feature extractor, we can efficiently enhance the model's performance.

3.3. Reprojection Refinement

To better leverage multi-view consistency as a constraint in the learning process, we implement a filtering mechanism to identify reliable surface points for accurate reprojection. Specifically, given reference and source views with known camera poses, we expect minimal discrepancies between non-occluded reference pixels and the reprojected pixels in neighboring views, both in terms of appearance and surface normals. Another hypothesis is that the predicted pixel depth should not undergo drastic changes across views during sequential movements.

Let $x \in \mathbb{R}^2$ be the 2D pixel coordinates and $\mathbf{X_w} \in \mathbb{R}^3$ be the world coordinates, We expect x can be reprojected to other views and contributes to its correlated features only. Then we define $\pi : \mathbb{R}^2 \mapsto \mathbb{R}^3$ as the projection function consist of camera intrinsic parameters $\mathbf{K}$ and camera pose $[\mathbf{R}|\mathbf{t}]$, which projects x to $\mathbf{X_w}$. Such that, reprojected pixel x_i' in the adjacent reference image can be computed by

$$x_i' = \pi'^{-1}(\pi(x_i)) \tag{7}$$

where π' denotes the projection of the reference view to the 3D world coordinates.

If the reprojected x_i' is within the image plane of the reference view, it is considered as a candidate reprojection point. We then render x_i' again to obtain its new normal and depth, denoted as $\mathbf{n}'$ and $\mathbf{d}'$, respectively. We expect the normals $\mathbf{n}$ and $\mathbf{n}'$ from the same surface point to be geometrically consistent in the world frame. Meanwhile, the depth values $\mathbf{d}$ and $\mathbf{d}'$ should not change significantly to avoid occluded pixels. To filter out outlier candidates, we use the L1 loss between $\mathbf{d}$ and $\mathbf{d}'$, and the cosine similarity between $\mathbf{n}$ and $\mathbf{n}'$. We consider points to be robust if their depth values and normal vectors both satisfy the following conditions, where $\mathbf{I}$ denotes the set of reference views, and $\mathbf{d_i}$ and $\mathbf{n_i}$ represent the depth value and normal vector of the i-th point, respectively:

$$\sum_{i}^{I} ||\mathbf{d_i} - \mathbf{d_i}'||_1 \leq \text{thresh}_1, \sum_{i}^{I} \frac{\mathbf{n_i} \cdot \mathbf{n_i}'}{||\mathbf{n_i} \cdot \mathbf{n_i}'||_2} \geq \text{thresh}_2. \tag{8}$$

where thresh_1 and thresh_2 are predefined thresholds.

3.4. Adaptive Correspondence Sampling

Existing research indicates that MLPs used for implicit neural representation are incapable of reconstructing signals with fine detail [33, 41]. Consequently, the learned geometry, such as SDF, tends to exhibit smoothness due to lack of detail. To address this issue, we propose an adaptive sampling strategy to increase the number of samples in the regions we wish to preserve fine detail.

In conventional image processing, handcrafted features [22] are often used to find matching pixels between images. Similarly, we look for pixel correspondences at the boundaries or regions with drastic changes in image gradients. We first search for a set of candidate correspondences $P_{\text{candidate}}$ across two adjacent frames with DFM [5]. Then we follow the outlier removal strategy introduced in section 3.3, but with only surface normal threshold. The set of final selected correspondences $\{(p_i, \hat{p}_i) \in P_{\text{final}}\}$ are considered to be stable matches. Subsequently, we formulate a loss term based on these extra sampled pixel correspondences. This enforces our model to assign greater weight to samples within the regions of our interest, particularly where finer details are located.

3.5. Optimization

Photometric Loss: 2D observations are the most direct way of learning to reconstruct the scene representation with multi-view consistency, which can be optimized via photometric reconstruction loss:

$$\mathcal{L}_{\text{photometric}} = \sum_{r \in \mathcal{R}} \|\hat{C}(r) - C(r)\|_1 \tag{9}$$

where $\mathcal{R}$ denotes the rays in each sampling batch, and C is the pixel color intensity.

Feature Loss: We adopt the MSE and L1 loss to impose surface feature consistency between the training frame and its adjacent reference frames. Here, $\hat{x}_i$ denotes any reprojected surface point that successfully passes our reprojection refinement stage, while $\bar{x}_i$ denotes its negative sample.

$$\begin{aligned}\mathcal{L}_{\text{feature}} = \frac{1}{n}\sum_{i=1}^{n}\Big(&\|f_\theta(x_i) - f_\theta(\hat{x}_i)\|_2 + \|f_\theta(x_i) - f_\theta(\hat{x}_i)\|_1 \\ &- \|f_\theta(x_i) - f_\theta(\bar{x}_i)\|_2 - \|f_\theta(x_i) - f_\theta(\bar{x}_i)\|_1\Big)\end{aligned} \tag{10}$$

where n is the number of sampled points remained after our reprojection refinement procedure.

Semantic Loss: The semantic loss in our algorithm consists of two terms, the first term is the pixel-wise cross-entropy loss between our predicted labels and pseudo ground-truth labels of the reference frame. The second term computes the pixel-wise cross-entropy loss between the pseudo ground-truth reference labels and the reprojection of predicted source labels in the reference view:

$$\begin{aligned}\mathcal{L}_{\text{semantic}} = &-\sum_{r \in \mathcal{R}}\sum_{c \in \mathcal{C}} q_{t,c}(r) \log \hat{q}_{t,c}(r) \\ &- \tau \sum_{r \in \mathcal{R}}\sum_{c \in \mathcal{C}} q_{t,c}(r) \log \hat{q}_{t+1,c}(r)\end{aligned} \tag{11}$$

where $\mathcal{C}$ represents the set of class labels, τ stands for warm-up ratio. $q_{t,c}$ and $\hat{q}_{t,c}$ refer to the ground truth and predicted semantic probability at class c of the reference frame, while $\hat{q}_{t+1,c}$ denotes the predicted semantic probability at class c of the source frame warped to the reference view.

Geometric loss: To ensure consistency between the volume-rendered normals $\hat{N}$ and the predicted monocular normal constraint $\bar{N}$, we follow the approach used in NeuRIS [47] and MonoSDF [56]. We minimize their L1 loss while maximizing their angular loss. Also, we followed the depth loss function from Midas [34]. In this approach, learnable scale and shift parameters, denoted as w and s, are utilized to align the model-predicted depth $\hat{D}$ with the monocular depth constraints $\bar{D}$.

$$\mathcal{L}_{\text{normal}} = \sum_{r \in \mathcal{R}} \|\hat{N}(r) - \bar{N}(r)\|_1 + \|1 - \hat{N}(r) \cdot \bar{N}(r)\|_1 \tag{12}$$

$$\mathcal{L}_{\text{depth}} = \sum_{r \in \mathcal{R}} \|(w\hat{D}(r) + s) - \bar{D}(r)\|_2 \tag{13}$$

Sampled Correspondence Loss: For the correspondences selected by our adaptive sampling process, we expect them to have similar surface normals and features. We optimize these correspondences based on their feature and normal loss:

$$\mathcal{L}_{\text{corr}} = \mathcal{L}_{\text{normal}}(p_i, \hat{p}_i) + \mathcal{L}_{\text{feature}}(p_i, \hat{p}_i) \tag{14}$$

Eikonal Loss: Same as the previous works followed IGR [10], we add an Eikonal term on sampled points to regularize SDF values in 3D space as well, where we denote $\mathcal{X}$ is a union set of uniformly sampled points and near-surface points.

$$\mathcal{L}_{\text{eikonal}} = \sum_{x \in \mathcal{X}} \left(\|\nabla f_\phi(x)\|_2 - 1 \right)^2 \tag{15}$$

Total Loss: Here, we define our total loss function as follows, note that the λ is only for balancing the magnitude because we have different objective functions for each sub-loss.

$$\begin{aligned} \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rgb}} + \lambda_1 \mathcal{L}_{\text{eikonal}} + \lambda_2 \mathcal{L}_{\text{depth}} + \lambda_3 \mathcal{L}_{\text{normal}} \\ + \lambda_4 \mathcal{L}_{\text{feature}} + \lambda_5 \mathcal{L}_{\text{semantic}} + \lambda_6 \mathcal{L}_{\text{corr}} \end{aligned} \tag{16}$$

4. Experiments

We have conducted excessive experiments in indoor scenes to assess the effectiveness of our proposed method. The experiments demonstrate that incorporating object surface features that satisfy multi-view consistency as guidance leads to improved 3D reconstruction results. Furthermore, a complementary relationship can be observed between our dynamic sampling and surface features. Our results achieve state-of-the-art performance on several datasets.

Datasets Our proposed method is evaluated on two datasets: ScanNet [4] and Replica [44]. ScanNet is a large-scale dataset that comprises 1613 indoor scenes, each accompanied by camera intrinsics and poses. The ground truth for 3D reconstruction in ScanNet is obtained by fusing RGBD camera data. In contrast, Replica is a meticulously synthetic virtual indoor scene dataset. It utilizes an accurate synthesis process to generate scenes with precise scene geometry and appearance, including camera intrinsics and poses.

Baselines We compare our method against various state-of-the-art techniques. For implicit representation models, we compare ours with other deep-learning approaches including UNISURF [30], VolSDF [55], Manhattan-SDF [11], NeuS [48], as well as the latest NeuRIS [47] and MonoSDF [56], which integrate monocular constraints within training phrase. For traditional approaches, we compare ours against COLMAP [36].

Evaluation Metrics: Following previous work [11, 15, 25, 27], we utilize several metrics to evaluate our results. For the Replica dataset, we report the F-score (using a threshold of 5 cm), Normal Consistency measure, and the Chamfer Distance. On the other hand, for the ScanNet dataset, our reporting metrics include the Chamfer Distance, Precision, Recall, and the F-score(with a threshold of 5 cm), as well as Accuracy and Completeness (details in the supplementary). It is worth noting that the F-score is often regarded as the most appropriate metric for assessing geometry quality, as argued in [45]. More details on our evaluation metrics can be found in the supplementary.

4.1. Experiment Results

ScanNet: Using the ScanNet dataset, we assessed the effectiveness of our proposed PMVC on real-world scenes. By ensuring multi-view consistency, our method not only recovers missing details but also improves overall smoothness, as seen in Figure S8. Compared to other methods, ours outperforms all of them as shown in Table 1. For full reports, please refer to the supplementary material.

Metric	Chamfer-L1 ↓	Prec↑	Recall↑	F-score↑
COLMAP [36]	0.141	71.1	44.1	53.7
UNISURF [30]	0.359	21.2	36.2	26.7
NeuS [48]	0.194	31.3	27.5	29.1
VolSDF [55]	0.267	32.1	39.4	34.6
Manhattan [11]	0.070	62.1	58.6	60.2
NeuRIS [47]	0.050	71.7	66.9	69.2
MonoSDF [56]	0.042	79.9	68.1	73.3
Ours	**0.038**	**81.5**	**77.4**	**79.4**

Table 1. Experiment results on ScanNet dataset. More discussion can be found in section 4.2

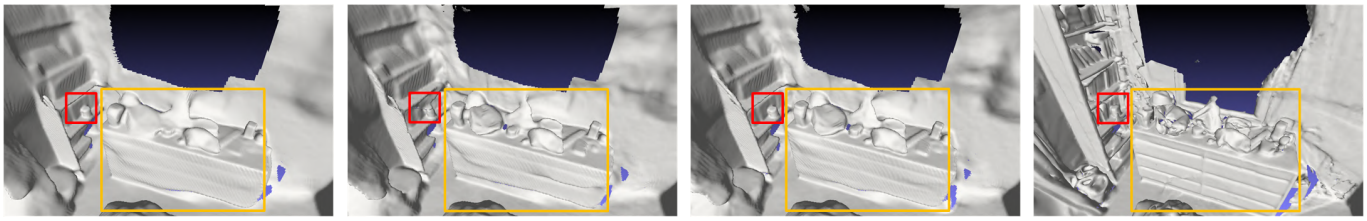
Replica: In addition to the Scannet, we also evaluate our method on synthetic indoor scenes from the Replica dataset. This analysis serves to demonstrate that adhering to multi-view consistency yields superior performance compared to current state-of-the-art methods as shown in Table 2.

Metric	Normal C.↑	Chamfer-L1 ↓	F-score ↑
VolSDF [55]	86.48	6.75	66.88
UNISURF [30]	90.96	4.93	78.99
MonoSDF [56]	92.11	2.94	86.18
Ours	**94.11**	**2.73**	**89.95**

Table 2. Experiment results on Replica dataset.

4.2. Ablation Study

We conducted an ablation study to assess the effectiveness of our multi-level constraints on the reconstruction quality on the Replica dataset [44]. We specifically isolated different priors, excluding all other constraints each time, to evaluate their individual contribution to the improvement of the reconstruction (Table 3). Our findings demonstrate that including these constraints effectively enhances the quality of the output. Notably, optimal performance was achieved by combining all types of constraints with adaptive sampling. Therefore, our model gains a better understanding of multi-view consistency, leading to improved reconstruction results. Here we also provide the visual result for the proposed adaptive sampling strategy.



(a) w/o adaptive sampling (b) w/$\mathcal{L}_{\text{corr}} = \mathcal{L}_{\text{normal}}$ (c) w/$\mathcal{L}_{\text{corr}} = \mathcal{L}_{\text{normal}} + \mathcal{L}_{\text{feature}}$ (d) Ground-Truth

Figure 4. This figure clearly demonstrates that our proposed adaptive sampling contributes to capturing fine details of objects in a scene. Additionally, we show that adaptive sampling integrates well with our multi-constraint approach.

w,w/o	Normal C.↑	Chamfer-L1 ↓	F-score ↑
baseline	92.11	2.94	86.18
+ corr	93.79	2.81	89.44
+ semantic	94.01	2.78	89.03
+ feature	94.02	2.82	89.60
Full Model	**94.11**	**2.73**	**89.95**

Table 3. Performance of different components of the proposed PMVC. From this table, we can observe the individual contributions of each component to the model's performance. The adaptive sampling (corr) improves the metrics, as does semantic and feature constraints. When all components are combined, the Full Model achieves the best performance.

Adaptive sampling visualization. In Figure 4, we visualize the ablation of adaptive sampling. The left image(Fig.4a) shows the reconstruction result without the adaptive sampling strategy, the center image (Fig.4b) shows the effect of adding adaptive sampling (normal loss only), while the image(Fig.4c) shows the result of the complete adaptive sampling (normal + feature loss). As observed from the center image, we can reconstruct more details of some smaller objects with higher accuracy. However, at this stage, we haven't incorporated our feature guidance, and it only relies on the normal loss $\mathcal{L}_{\text{corr}} = \mathcal{L}_{\text{normal}}(p_i, \hat{p}_i)$. Experiments have shown that the model performs better when having both normal and feature consistencies (see areas bounded in yellow). For instance, solely utilizing the normal loss from sampled correspondences tends to introduce excessive noise, resulting in artifacts in the reconstruction (see areas bounded in red), and this issue is mitigated after introducing the feature loss.

Robustness under Different View-angle. To assess the robustness of our model in maintaining good performance given varying view angles, we conducted experiments using the 10th frame (with major overlap), the 30th frame (with partial overlap), and the 50th frame (with minor overlap). All of these frames serve as reference views relative to the current frame. As shown in Table 4, our model can tolerate such view angles to a certain extent. Optimal results are achieved when the camera angles do not significantly differ.

Metric	Chamfer-L1 ↓	Prec↑	Recall↑	F-score↑
scene0050$_{w/o}$	0.046	72.4	64.6	68.3
scene0050$_{50th}$	0.042	78.9	72.5	75.6
scene0050$_{30th}$	**0.041**	**79.6**	72.8	76.0
scene0050$_{10th}$	**0.041**	**79.6**	**73.2**	**76.3**

Table 4. Performance of the proposed model varies minimally under different reference view-angle disparities. The results suggest that our model is robust to sparse reference views.

5. Conclusion

In this paper, we propose a novel multilevel prior strategy, PMVC, which emphasizes stable guidance under the principle of multi-view consistency. Our experiments show that providing models with stable priors while adhering to multi-view consistency can enhance their understanding of the scene, resulting in improved reconstruction outcomes. Moreover, our approach reduces the dependency on specific models. For instance, the addition of a simple CNN model can surpass current state-of-the-art techniques that rely on deeper networks. Furthermore, our ablation experiments demonstrate that the interplay between multi-level priors is complementary, and they can reciprocally constrain each other to achieve more stable scene reconstruction.

Limitations: Our method predominantly relies on specific monocular constraints, such as normal and depth maps. These maps act as geometric constraints, particularly crucial in areas with weak texture. Consequently, if these pretrained models are trained predominantly on indoor scenes, they may not perform optimally in outdoor scenes, leading to potential inaccuracies in predicting normal and depth maps. Another limitation is the pace of convergence. Due to an additional rendering process for pixels that can be reprojected to the reference view, our approach may take longer to train.

References

[1] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022. 2, 3

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 3

[3] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6970–6981, 2020. 3

[4] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3, 7

[5] Ufuk Efe, Kutalmis Gokalp Ince, and Aydin Alatan. Dfm: A performance baseline for deep feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4284–4293, June 2021. 6

[6] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022. 2, 3

[7] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. 3

[8] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 1, 3

[9] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 1

[10] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020. 7

[11] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5511–5520, 2022. 3, 7, 12, 14

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[13] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018. 3

[14] Hanqi Jiang, Cheng Zeng, Runnan Chen, Shuai Liang, Yinhe Han, Yichao Gao, and Conglin Wang. Depth-neus: Neural implicit surfaces learning for multi-view reconstruction based on depth information optimization. *arXiv preprint arXiv:2303.17088*, 2023. 3

[15] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18974, 2022. 7

[16] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 3

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 12

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 5, 12

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5

[20] Patrick Labatut, Jean-Philippe Pons, and Renaud Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007. 3

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. 2023. 5, 12

[22] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 6

[23] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 3

[24] N. Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 4

[25] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 7

[26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2, 3, 4

[27] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *ECCV*, 2020. 7

[28] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5

[29] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1, 3

[30] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 7, 14

[31] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3, 4

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 12

[33] Sameera Ramasinghe and Simon Lucey. Beyond periodicity: Towards a unifying framework for activations in coordinate-mlps. In *European Conference on Computer Vision*, pages 142–158. Springer, 2022. 6

[34] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. 6

[35] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 3

[36] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 7, 14

[37] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006. 3

[38] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023. 5

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 12

[40] Ayan Sinha, Asim Unmesh, Qixing Huang, and Karthik Ramani. Surfnet: Generating 3d shape surfaces using deep residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6040–6049, 2017. 2

[41] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 2, 6

[42] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3

[43] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 3

[44] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3, 7

[45] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. *CVPR*, 2021. 7

[46] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Itermvs: iterative probability estimation for efficient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8606–8615, 2022. 3

[47] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 139–155. Springer, 2022. 2, 3, 4, 6, 7, 12, 14

[48] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1, 3, 7, 14

[49] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 3

[50] Yusen Wang, Zongcheng Li, Yu Jiang, Kaixuan Zhou, Tuo Cao, Yanping Fu, and Chunxia Xiao. Neuralroom:

Geometry-constrained neural implicit surfaces for indoor scene reconstruction. *arXiv preprint arXiv:2210.06853*, 2022. 2

[51] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5483–5492, 2019. 1

[52] Fengting Yang and Zihan Zhou. Recovering 3d planes from a single image via convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 3

[53] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 3

[54] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4805–4815. Curran Associates, Inc., 2021. 3, 4

[55] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 7, 14

[56] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022. 1, 2, 3, 4, 6, 7, 12, 14

[57] Jingyang Zhang, Yao Yao, and Long Quan. Learning signed distance field for multi-view surface reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6525–6534, 2021. 3

[58] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 3

[59] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 3, 5

Supplementary Material for PMVC

In this supplementary document, we discuss the architectural and implementation details in Section 6. Next, in Section 7, we provide additional quantitative and qualitative results across the various datasets we experimented with, as well as our scene reconstruction results. Finally, we discuss the potential negative impact of this work in Section 8.

6. Implementation Details

In this section, we first describe additional details regarding our parameterizations and optimization in Section 6.1. Next, we present an overview of two different architectures for Feature Extractor in Section 6.2 and provide details of the semantic Cues in Section 6.3, and discuss evaluation metrics in Section 6.4.

6.1. Parameterizations

In this experiment, we utilized PyTorch [32] as our experimental framework and employed Adam [17] as our optimizer. Our code was initially set up using the framework from MonoSDF [56], and we adhered to their learning rate 5e-4. Empirically, we established λ_1, λ_2, λ_3, λ_4, λ_5 and γ at 0.1, 0.1, 0.05, 0.5, 0.04 and 0.01 respectively. For feature extraction, we utilized VGG-19 [39] model. All experiments were executed on a single RTX3090 GPU. To align with our task requirements, we make minor modifications to the NYU labels, reducing them to 38 classes representing indoor scene objects and eliminating ambiguous labels such as "other furniture". For additional details on semantic map generation, please refer to section 6.3.

Implement details. We implement our Neural Implicit Representation architecture using two MLPs. Each MLP corresponds to a 256-dimension feature. Our Neural Renderer network outputs both an RGB intensity and a label. As previously mentioned in the paper, we utilize MonoSDF for geometry initialization. In addition to this, we deploy our pipeline to generate a semantic map for each frame during the pre-processing phase 6.3. We optimize our model over 200,000 iterations. In terms of computational time, optimizing a single scene using the full model on a single NVIDIA RTX 3090 GPU takes around 24 hours. The feature constraint process requires approximately 18 hours. Meanwhile, adaptive sampling and semantic constraint take around 15 hours and 12 hours, respectively.

6.2. Feature extractors

In this section, we present results obtained from two distinct feature extractors, namely ResNet-18 and VGG-19. We carried out the ablation experiments on ScanNet, and the data presented in Table S5 discovers the differences between feature extractors. Additionally, it reveals a consistent improvement across the F-score.

Scene	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑	Chamfer-dist↓
MonoSDF	**0.035**	0.048	**79.9**	68.1	73.3	0.042
+ ResNet-18	0.041	0.043	76.4	72.5	74.3	0.042
+ VGG-19	0.040	**0.042**	79.8	**74.9**	**77.4**	**0.041**

Table S5. Two Feature Extractors comparison

6.3. Semantic Map Pipeline

In this section, we introduce our new semantic generation pipeline, as shown in Figure S5, which outperforms other methods. We also conduct several studies to verify the effectiveness of our method. These include comparing with the Manhattan fine-tuned model [11] (as presented in Table S6) and integrating our pipeline into the NeuRIS [47] framework (Table S7). We evaluate our semantic priors on ScanNet.

Thanks to the powerful zero-shot capability of the SAM model [18], we can perform pixel-level segmentation of arbitrary scenes. However, SAM does not possess label classes, which is a problem we sought to address. To rectify this, we first employed the method outlined in [21] for each image, then used prompt text hints to generate bounding boxes (BBox). Subsequently, we used SAM to execute pixel clustering segmentation within each BBox. To better adapt to our task, we eliminated semantically ambiguous portions of the NYU label, such as 'other furniture' and 'other structures'. As a result, our final label classes consist of 38 distinct categories. For the objects that are not included in the 38 classes, we define them as 'unknown', which will not contribute to our semantic optimization.

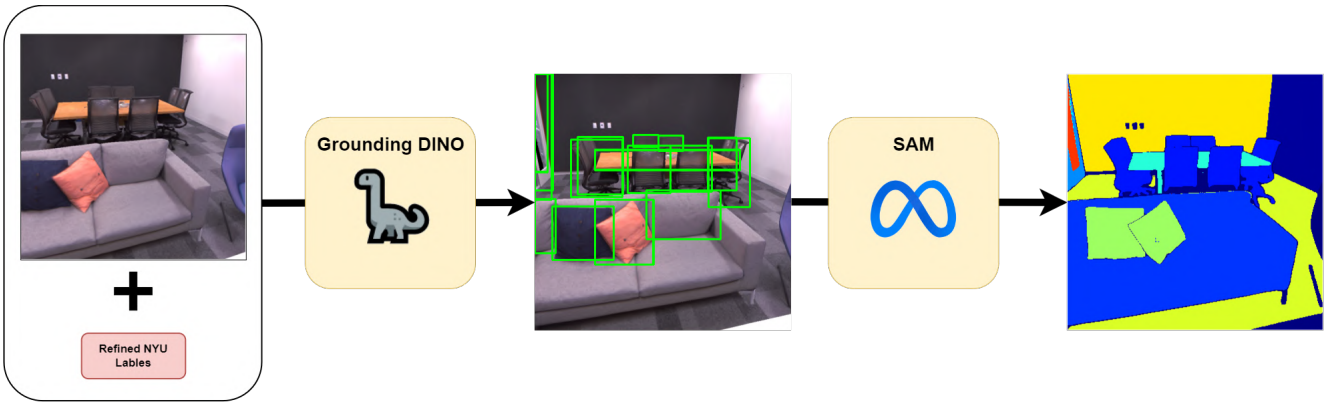


Figure S5. In this pipeline, we utilize the concept of Grounding DINO and incorporate modified NYU labels as our textual cues. Initially, we employ these cues to generate bounding boxes for each class. Subsequently, we apply the SAM (Segment Anything Model) to obtain pixel-wise semantic maps. This approach allows us to effectively map the textual information to the corresponding regions in the image, providing a detailed and accurate representation of the semantic content.

Model	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑	Chamfer-dist↓
MonoSDF	**0.035**	0.048	**79.9**	68.1	73.3	0.042
+ semantic cues (DeepLabV3)	0.040	0.042	77.0	73.5	75.1	0.041
+ semantic cues (Ours)	0.040	**0.041**	79.2	**75.6**	**77.3**	**0.040**

Table S6. Two pipeline comparison

Model	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑	Chamfer-dist↓
NeuRIS	0.050	0.049	71.7	66.9	69.2	0.050
+ semantic cues (DeepLabV3)	0.048	0.048	72.7	67.7	70.1	0.048
+ semantic cues (Ours)	**0.044**	**0.047**	**75.7**	**69.9**	**72.7**	**0.046**

Table S7. Semantic constraints experiments on NeuRIS

6.4. Evaluation Metrics

In line with prior research, we adopted several evaluation metrics to assess the quality of our reconstruction. For the ScanNet dataset, our report features Accuracy, Completeness, Chamfer Distance, Precision, Recall, and F-score. In contrast, for the Replica dataset, we present the Normal Consistency, Chamfer Distance, and F-score. Detailed definitions of these evaluation metrics are specifically provided in Table S8.

7. Additional Results

This section provides more qualitative and quantitative comparison results for the Replica (Figure S8) and ScanNet (Figure S7) datasets. In addition, we demonstrate the full evaluation metrics mentioned in the main paper (Table S9) and discuss the lack of a significant difference between the ACC scores of previous methods and our full model on ScanNet. Lastly, we provide some rendered images to show that our methods reduce the dependency on the performance of pre-trained models (Figure S9).

Performance discussion. Upon closer examination (Figure S6), we found that the ground truth provided by ScanNet tends to be quite noisy. For instance, we observed missing objects that should be present in the ground truth (as illustrated in Supplementary Figure S5). This noise and incompleteness in the ground truth might affect the evaluation metrics, potentially leading to the observed inconsistent improvement in the ACC performance between our result (0.038) and MonoSDF's result (0.035). Notably, this issue does not arise in the synthetic dataset, Replica.

Metric	Definition
Acc	$\underset{p \in P}{\text{mean}}(\min_{p^* \in P^*} ||p - p^*||_1)$
Comp	$\underset{p^* \in P^*}{\text{mean}}(\min_{p \in P} ||p - p^*||_1)$
Chamfer	$\frac{\text{Acc}+\text{Comp}}{2}$
Precision	$\underset{p \in P}{\text{mean}}(\min_{p^* \in P^*} ||p - p^*||_1 < 0.05)$
Recall	$\underset{p^* \in P^*}{\text{mean}}(\min_{p \in P} ||p - p^*||_1 < 0.05)$
F-score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
Normal-Acc	$\underset{p \in P}{\text{mean}}(n_p^T n_{p*})$ s.t. $p^* = \underset{p^* \in P^*}{\text{argmin}} ||p - p^*||_1$
Normal-Comp	$\underset{p^* \in P^*}{\text{mean}}(n_p^T n_{p*})$ s.t. $p = \underset{p \in P}{\text{argmin}} ||p - p^*||_1$
Normal-Consistency	$\frac{\text{Normal-Acc}+\text{Normal-Comp}}{2}$

Table S8. Evaluation Metrics. We present the evaluation metrics along with their definition, which we employ to assess the quality of reconstruction. P and P^* represent the point clouds obtained from the predicted and the actual mesh, respectively. n_p stands for the normal vector at the point p.

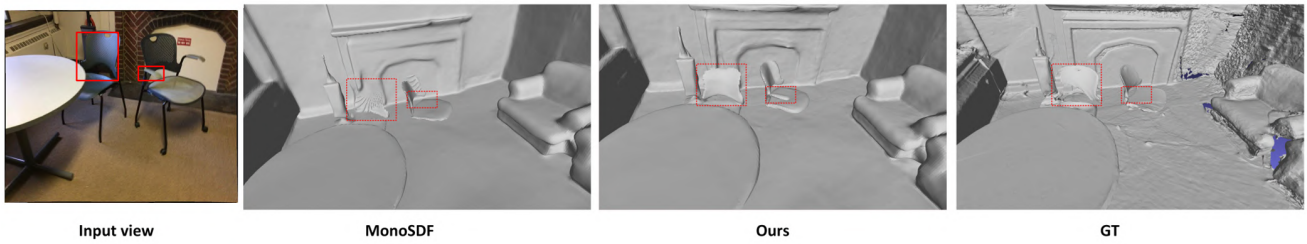


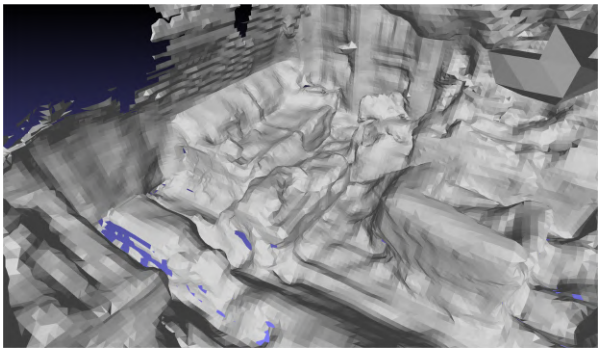
Figure S6. By examining the input views, it is evident that the chair possesses handles and a complete back part. Surprisingly, this crucial information is missing in the provided ground truth (GT). In contrast, our proposed method excels in reconstructing this missing part, highlighting the capability of our approach in capturing and reproducing fine details accurately.

Metric	Acc↓	Comp↓	Chamfer-L1 ↓	Prec↑	Recall↑	F-score↑
COLMAP [36]	0.047	0.235	0.141	71.1	44.1	53.7
UNISURF [30]	0.554	0.164	0.359	21.2	36.2	26.7
NeuS [48]	0.179	0.208	0.194	31.3	27.5	29.1
VolSDF [55]	0.414	0.120	0.267	32.1	39.4	34.6
Manhattan [11]	0.072	0.068	0.070	62.1	58.6	60.2
NeuRIS [47]	0.050	0.049	0.050	71.7	66.9	69.2
MonoSDF [56]	**0.035**	0.048	0.042	79.9	68.1	73.3
Ours	0.038	**0.039**	**0.038**	**81.5**	**77.4**	**79.4**

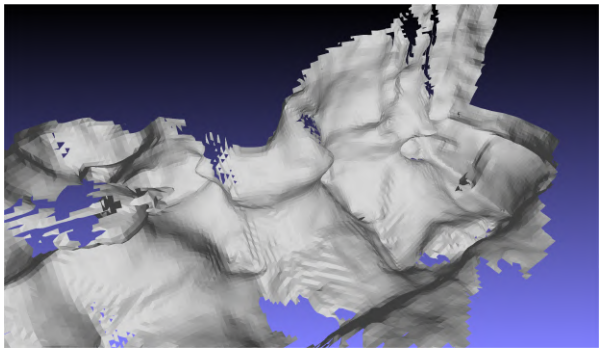
Table S9. Full results on ScanNet dataset.

8. Societal Impact

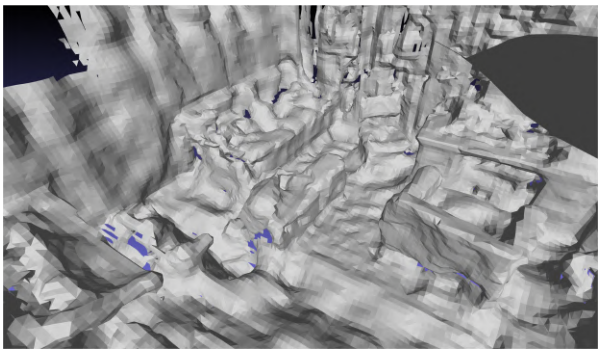
Our proposed method has the potential for significant improvements in 3D reconstruction from multiple viewpoints, which can be applied to virtual reality or greatly reduce the modeling time for designers. However, there are some drawbacks to consider. One drawback is that our approach requires relatively dense inputs; otherwise, multi-view consistency is hard to obtain. Besides, we did not impose additional constraints on the reconstruction process, which may raise privacy concerns when applied to indoor scene reconstruction. Additionally, our training time is relatively long, resulting in increased power consumption. However, further engineering improvements may address this environmental issue in the future.



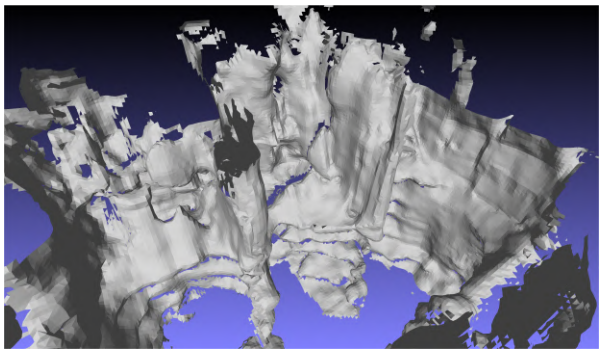
(a) NeuS, Scene 1



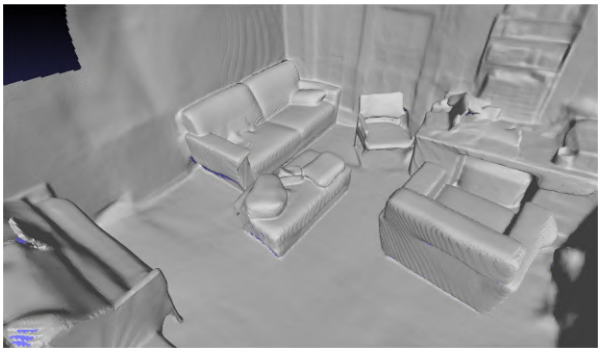
(b) NeuS, Scene 2



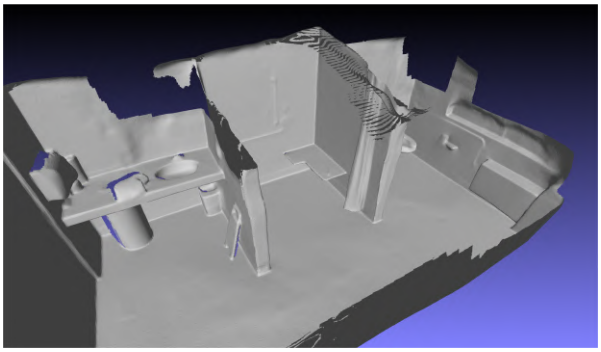
(c) Volsdf, Scene 1



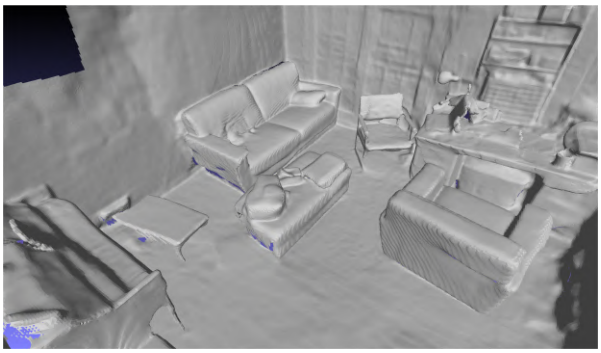
(d) Volsdf, Scene 2



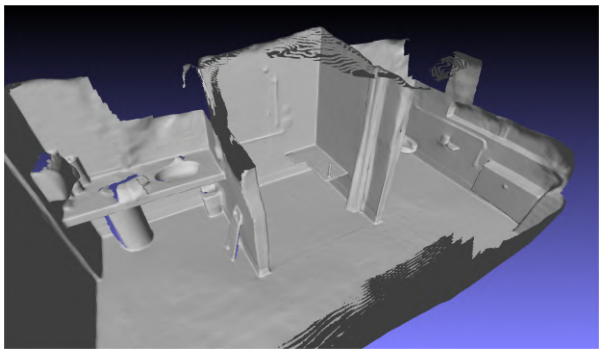
(e) Monosdf, Scene 1



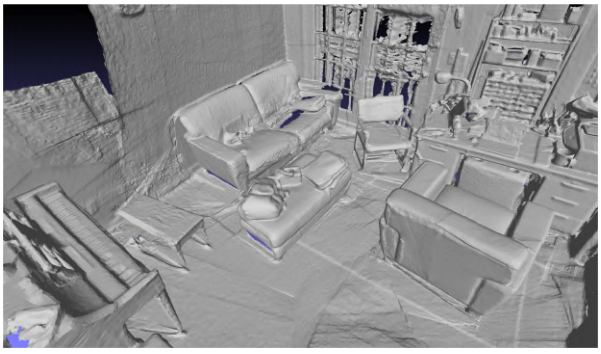
(f) Monosdf, Scene 2



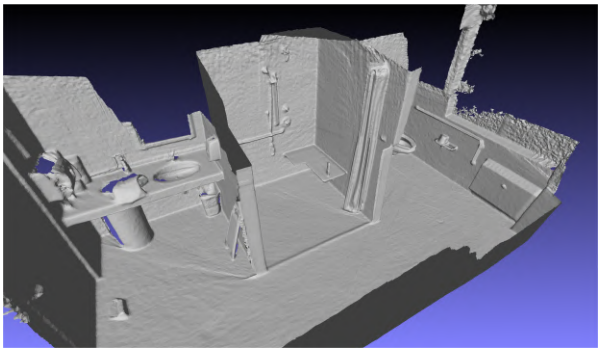
(g) Ours, Scene 1



(h) Ours, Scene 2



(i) ground truth, Scene 1



(j) ground truth, Scene 2

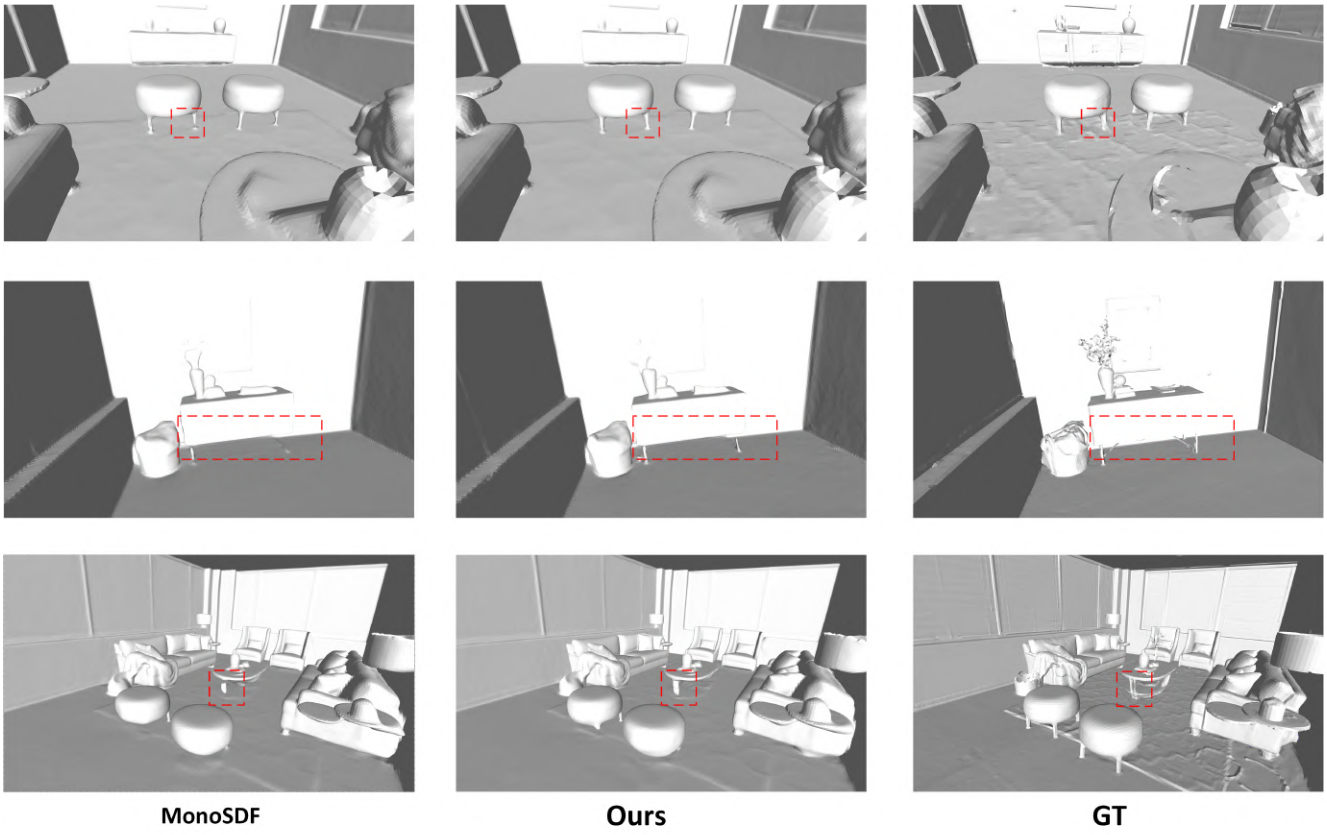
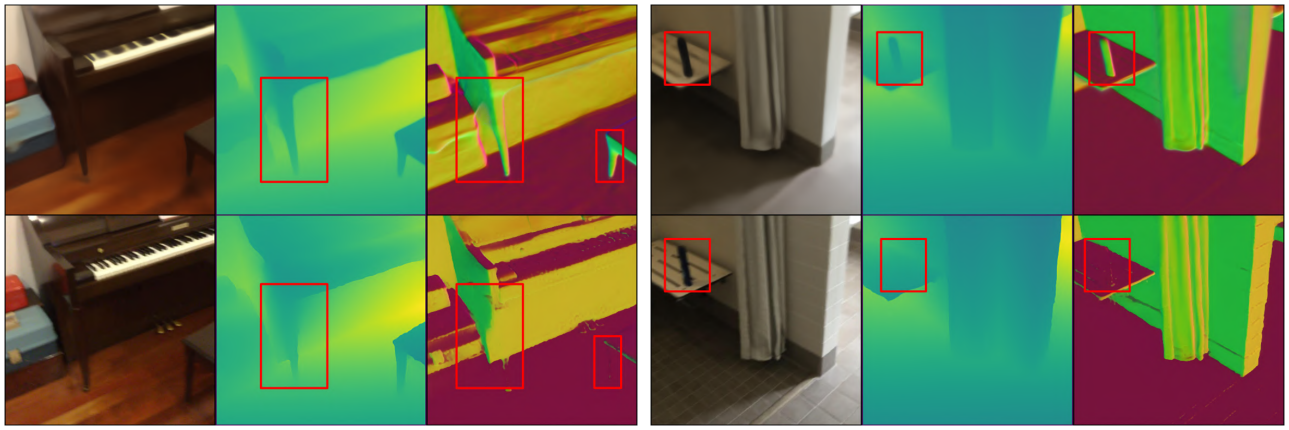


Figure S8. We compare our proposed technique with MonoSDF and the ground truth. As highlighted by the rectangles, our technique shows improvements. We can observe that, compared to previous methods that are only using pre-trained models, our technique reconstructs fine detail well.



(a) case1: pseudo ground truth shows inconsistency

(b) case2: pseudo ground truth provides negative guidance

Figure S9. Rendered results from the ScanNet dataset are presented. The second row shows the RGB image used as our input. The depth map and normal map were estimated using pre-trained models. We observed that the pseudo-ground truth does not always help the model understand scenes due to the potential limitations of the pre-trained models. In contrast, our rendering results exhibit superior details that help mitigate this issue.